Balanced Cluster Canonical Correlation Analysis for Multimodal Data Integration in Alzheimer's Disease

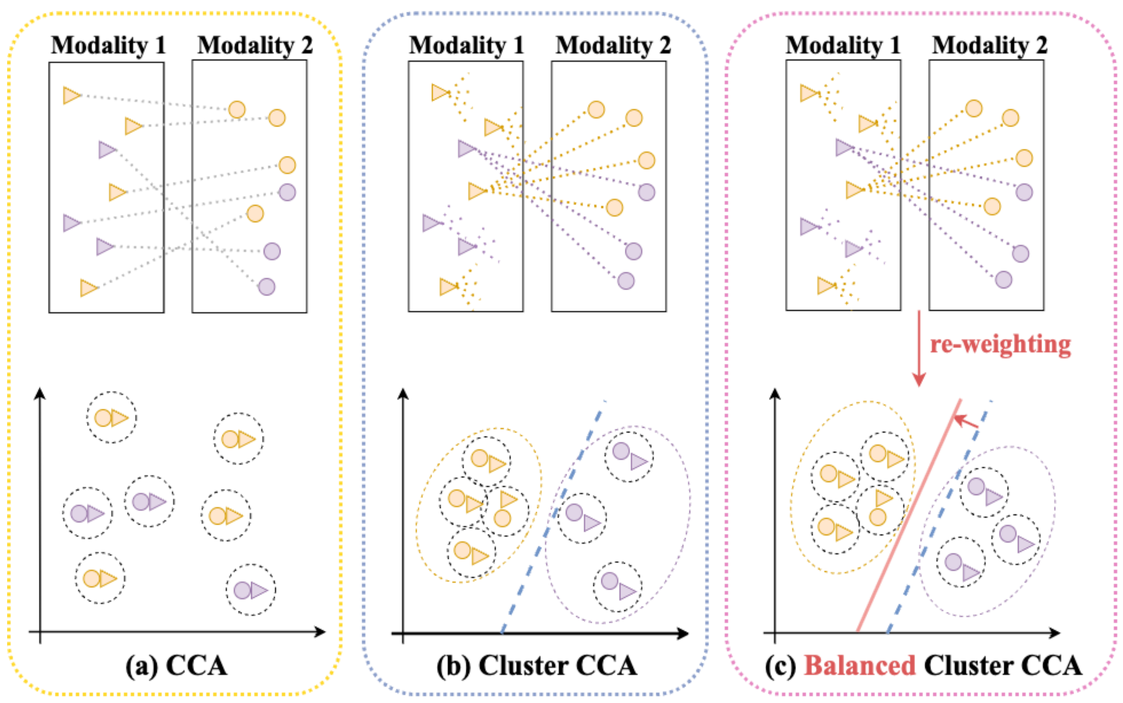
Boning Tong[1*], Zhuoping Zhou[1*], Jingxuan Bao[1], Bojian Hou[1], Jason H. Moore[2], Andrew J. Saykin[3], Marylyn D Ritchie[1], Qi Long[1], Li Shen[1], for the ADNI and the ADSP

[1]University of Pennsylvania, Philadelphia, PA, USA; [2]Cedars Sinai Medical Center, West Hollywood, CA, USA;
[3]Indiana University, Indianapolis, IN, USA;
*These authors contributed equally to this work. † Correspondence to Li Shen (li.shen@pennmedicine.upenn.edu)

Introduction

- **Canonical Correlation Analysis (CCA)** is a multivariate statistical technique that find linear combinations of variables from two sets that **maximize the correlation** between them, **reveal underlying patterns** and shared structures.
- The applications of CCA span a wide range of fields, each demonstrating its effectiveness in **handling multidimensional data**, especially in Alzheimer's Disease (AD) with multiple imaging modalities or imaging genomics analysis.
- **Problem:** The **label imbalance** in AD cohorts results in the **underrepresentation of minority AD** subjects. Such imbalance skews model learning, producing **biased CCA representations** dominated by the majority group. Consequently, this distortion limits model reliability, delays early diagnosis and precise intervention.
- Existing solutions to imbalance issue including using oversampling or cost-sensitive learning in combination with CCA. However, they will modify the original data distribution or rely on adjustments in subsequent classification models.
- **Goal:** propose an approach that **directly learns a balanced representation** within CCA for the class imbalanced dataset.

Methods



$$\max_{\mathbf{U},\mathbf{V}} \ \text{trace}\left(\mathbf{U}^\top \mathbf{X}^\top \mathbf{Y}\mathbf{V}\right)$$
$$\text{subject to} \quad \mathbf{U}^\top \mathbf{X}^\top \mathbf{X}\mathbf{U} = \mathbf{V}^\top \mathbf{Y}^\top \mathbf{Y}\mathbf{V} = \mathbf{I}_r.$$

Traditional CCA identifies **pairwise correspondences**.

$$\max \ \text{trace}\left(\mathbf{U}^\top \bar{\mathbf{\Sigma}}_{\mathbf{xy}} \mathbf{V}\right)$$
$$\text{subject to} \quad \mathbf{U}^\top \bar{\mathbf{\Sigma}}_{\mathbf{xx}} \mathbf{U} = \mathbf{V}^\top \bar{\mathbf{\Sigma}}_{\mathbf{yy}} \mathbf{V} = \mathbf{I}_r,$$

Cluster CCA enhances **within-class correspondences**, improving class **separability**.

$$\bar{\mathbf{\Sigma}}_{\mathbf{xx}} := \frac{1}{\sum_{c=1}^{C} |\mathbf{X}^c||\mathbf{Y}^c|} \sum_{c=1}^{C} \frac{\sum_{c'=1}^{C} n_{c'}}{n_c} \sum_{j=1}^{|\mathbf{X}^c|} |\mathbf{Y}^c| \mathbf{x}_j^{c\top} \mathbf{x}_j^c,$$
$$\bar{\mathbf{\Sigma}}_{\mathbf{yy}} := \frac{1}{\sum_{c=1}^{C} |\mathbf{X}^c||\mathbf{Y}^c|} \sum_{c=1}^{C} \frac{\sum_{c'=1}^{C} n_{c'}}{n_c} \sum_{k=1}^{|\mathbf{Y}^c|} |\mathbf{X}^c| \mathbf{y}_k^{c\top} \mathbf{y}_k^c,$$
$$\bar{\mathbf{\Sigma}}_{\mathbf{xy}} := \frac{1}{\sum_{c=1}^{C} |\mathbf{X}^c||\mathbf{Y}^c|} \sum_{c=1}^{C} \frac{\sum_{c'=1}^{C} n_{c'}}{n_c} \sum_{j=1}^{|\mathbf{X}^c|} \sum_{k=1}^{|\mathbf{Y}^c|} \mathbf{x}_j^{c\top} \mathbf{y}_k^c.$$

Balanced Cluster CCA employs **re-weighting** to further **balance the representations** of two classes, widening the **margin for minority** classes.

ADSP Data

Modality	# features	Diagnosis Distribution	Canonical Dimension
MRI Volume	151	977 CN, 512 MCI, 322 AD	10
SNPs	1789		

- 1,811 subjects from multiple cohorts with their first visit (ADRC, WHICAP, WRAP, and ADNI at the R4 phase)
- Binary classifications: CN vs. AD, CN vs. MCI and MCI vs. AD
- Models:
  - CCA, Cluster CCA, Balanced Cluster CCA
  - CCA models + SVM, original data + SVM

Results

❖ AD Classification Performance with ADSP data

- For the **high-dimensional** ADSP data, even the **traditional CCA+SVM** method achieves **better** predictive performance than **directly applying SVM** to the original data, highlighting the advantage of CCA in handling high-dimensional data.
- Our proposed **Balanced Cluster CCA** consistently achieves the **highest balanced accuracy** across all classification tasks, demonstrating its effectiveness in representing imbalanced-label data and improving classification performance in Alzheimer's disease studies.

ADSP IMAGE AND SNPS DATASETS.

Methods	CN vs. AD		CN vs. MCI		MCI vs. AD	
	BACC	ACC	**BACC**	ACC	**BACC**	ACC
SVM	0.784±0.041	0.850±0.026	0.651±0.017	0.704±0.020	0.740±0.020	0.759±0.015
CCA+SVM	0.797±0.029	0.872±0.019	0.676±0.039	0.729±0.032	0.764±0.029	0.784±0.022
C-CCA+SVM	0.821±0.011	0.888±0.011	0.703±0.046	0.739±0.045	0.751±0.042	0.777±0.034
BC-CCA+SVM	**0.841±0.014**	0.894±0.010	**0.730±0.018**	0.761±0.015	**0.778±0.038**	0.796±0.030

Discussion

- Produces **balanced** cluster densities in the low-dimensional **representations** by tightening minority class representation, relaxing compactness in majority class, and enhances between-class separation
- **Future directions:**
  - Develop non-linear extensions
  - Integrate other multi-view learning approaches
  - Broader downstream tasks across various fields and datasets

Conclusion

The Balanced Cluster CCA effectively **addresses class imbalance** in Alzheimer's disease research through reweighting. It **outperforms standard CCA** methods in classification tasks and shows strong potential for broader biomedical applications involving imbalanced data.

Acknowledgements

This work was supported in part by the NIH grants R01 LM013463, U01 AG066833, U01 AG068057, P30 AG073105, R01 AG071470, U19 AG074879 (CLEAR-AD), R01 AG071174, RF1 AG063481, U19 AG024904 (ADNI), U24 AG074855 (ADSP PHC), and P30 AG072976 (Indiana ADRC). Data were obtained from the Alzheimer's Disease Neuroimaging Initiative and Alzheimer's Disease Sequencing Project database.